Optimal test and sampling designs for polytomous item response theory models

Summary

The objective of this thesis is to apply statistical methods of design optimisation to improve estimation of parameters of polytomous Item Response Theory (IRT) models. For these models, no global optimal designs can be found, i.e. optimal, irrespective of the values the model parameters might take. The reason for that lies in the dependence of non-linear models' information matrices on the unknown parameter values. Local optimality, i.e. designs that are optimal for a given set of parameters, can be found but usually offers a spurious solution to the design problem. Throughout this thesis a pre-specification of a parameters' range is used to handle the lack of knowledge of the parameters values.

Chapter one is introductory. It describes two polytomous IRT models to be dealt with in this thesis: the Nominal Response Model (NRM) and the Graded Response Model (GRM). The core of statistical techniques of design optimisation within the context of Item Response Theory (IRT) is shortly presented. In IRT, the issue of design optimisation is usually differentiated in two cases: optimal sampling or calibration design and optimal test design. In the former an optimal sample of test-takers is sought over a parameters space of latent trait values to efficiently estimate the item parameters. In optimal test design, which aims at a precise and accurate estimation of multiple or single trait values by selecting the set of most informative items from an item pool (item parameters space), a further differentiation between group-based and individual trait estimation can be made.

Chapter two deals with optimal calibration of items described by the NRM. An approach based on the concept of relative efficiency within a maximin framework is implemented by means of the D-optimality criterion. By varying the composition of the parameter space, it is shown that the higher the parametric diversity of the item space, the greater the efficiency loss of the maximin design. Alternative designs, like designs with a uniform and normal distribution of latent trait values also display this effect. With relative efficiencies often comparable to the ones of maximin designs, the uniform distribution emerges as a robust design against efficiency loss. Maximin designs offer a quantitative basis, upon which solid comparisons among different sampling designs can be drawn.

Chapter three uses the optimal design approach to select NRM items for a group-based, fixed-form test. Two objective functions of Fisher's Information, the A- and D-optimality criteria are applied and their selective strategies investigated. The criteria final test functions are evaluated as a function of varied latent trait distributions and item pool compositions. The distinct test assembly by the different criteria raises the possibility of

matching the item selection criterion to the specific test goals, usually expressed in form of a target information function. The obtained results suggest a greater sensitivity of the D-optimality criterion to changes of the underlying trait distribution. The A-criterion, by contrast, seems to manifest a clear preference for highly discriminative items, irrespective of the trait distribution. Due to the more accentuated dependence on the trait distribution of the item selection by the D-criterion, its suitability for computerised adaptive testing (CAT) is anticipated.

Chapter four evaluates the performance of the D-optimality criterion in CAT, demonstrating its feasibility for individually tailored tests. Two other criteria, A-optimality the Kullback-Leibler (KL) information criteria, are also considered for comparative purposes. A CAT simulation study is carried out, giving special emphasis to the initial phase of trait estimation. The so-called CAT early stage can be characterised by instability of the trait estimates. It is known that the estimates in this stage need to be corrected for bias. In general, this study shows that the spread of item bank information along the latent trait continuum is a major factor determining accuracy, precision and stability of trait estimation. The more uniform this information spread, the greater the quality of estimation for all applied criteria. For less uniformly spread information, all criteria perform similarly in terms of measurement accuracy and precision at the end of the test. They, however, differ in estimation stability during the early stage. The A-criterion displayed accentuated fluctuations of mean squared errors (MSE), while the D- and KL-criteria seemed to circumvent the attenuation paradox more effectively. Due to its flexibility of item selection, switching between low and high discriminative items as the test unfolds, it is conjectured that the D-criterion could make a more balanced use of the available items in a pool.

Chapter five extends the same methods from chapter four to an adaptive test with the GRM with reproducible results, including the robustness of the D-optimality criterion against early stage instability. The criteria preferences with respect to the parameter characteristics of the selected items are more closely investigated as an underpinning explanation for the criteria performances. The additional issue of item exposure is also evaluated. No measure for controlling the criteria item exposure rates is applied. The foregoing assumption as regards D-criterion's balanced use of items in a pool (chapter 4) could not be fully confirmed. In the matter of item pool usage, the larger the pool (with higher parametric diversity) the more overlapping the exposure rates of the three applied criteria.

In literature references there is a lack of congruity among opinions as regards which sort of items, polytomous or dichotomous, supply more information on the unknown latent trait. Generally, it is taken for granted that polytomous items are more informative than

dichotomous. While this assertion is true for polytomous items described by the GRM, its validity is questionable for items with a nominal response format. Chapter six shows through a polytomous-dichotomous transition of the NRM that, given due consideration to item discrimination parameters, dichotomous items can be as informative as their polytomous counterpart, if not more. The findings of test design optimisation, in which items are selected from a composite pool with polytomous NRM and dichotomous 2PL items, supports this statement and shows that the general view favouring polytomous items is not always justified for the NRM. The test' target information function and the test's setting (paper and pencil or adaptive) are major factors to be considered in the choice between dichotomous and polytomous items. It is argued that the gain in measurement precision, that might be achieved through polytomous items, could be traded-off for the parametric simplicity of dichotomous items.

Samenvatting

Het doel van dit proefschrift is om statistische methoden voor design optimalisatie te gebruiken voor efficiënte schatting van parameters in polytome Item-Respons-Theorie (IRT) modellen. Voor deze modellen kunnen geen globale optimal designs worden gevonden, d.w.z. optimaal voor alle mogelijke waarden die de model parameters zouden kunnen aannemen. De reden daarvoor ligt in de afhankelijkheid van de informatiematrices bij niet-lineaire modellen van de onbekende parameters. Lokale optimaliteit, d.w.z. designs die optimaal zijn voor een gegeven set van parameter waarden, kunnen gevonden worden maar bieden gewoonlijk een niet juiste oplossing voor het design probleem. In dit proefschrift wordt voornamelijk aan de hand van een gespecificeerde design de informatie over de parameters gemaximaliseerd.

Hoofdstuk een is inleidend. Het beschrijft de twee polytome IRT modellen waarmee in dit proefschrift gewerkt wordt: het Nominale Responsmodel (NRM) en het Graded Responsmodel (GRM). De belangrijkste statistische technieken voor design optimalisatie binnen de context van de Item-Respons-Theorie (IRT) worden in het kort gepresenteerd. In IRT wordt design optimalisatie gewoonlijk onderscheiden in twee gevallen: optimale steekproef of calibratie design en optimale test design. In het eerste geval wordt een optimale steekproef van test-nemers gezocht om de item parameters zo efficiënt mogelijk te schatten. In optimaal test design, dat zich richt op een nauwkeurige en accurate schatting van de latente kenmerken door de meest informatieve verzameling van items uit een item 'pool' (item parameter ruimte) te selecteren, kan een verdere differentiatie gemaakt worden tussen groepgewijze en individuele schattingen van latente kenmerken.

Hoofdstuk twee behandelt optimale calibratie van items beschreven door het NRM. Een benadering gebaseerd op het concept van relatieve efficiëntie binnen een maximin kader d.m.v. het D-optimaliteitscriterium wordt toegepast. Door de samenstelling van de parameter ruimte te variëren wordt aangetoond dat hoe groter de spreiding van de items parameter waarden in de item-ruimte is, des te groter het verlies aan efficiëntie van het maximin design. Ook alternatieve designs, met een uniforme en normale verdeling van de waarden van de latente kenmerken vertonen dit effect. Het uniforme design heeft vaak vergelijkbare relatieve efficiëncies als het maximin design. Maximin designs bieden een kwantitatieve basis voor een zinvolle vergelijking tussen verschillende calibratiedesigns.

Hoofdstuk drie gebruikt de optimal design benadering om NRM items te kiezen voor een 'fixed-form' toets. Twee doelfuncties van Fisher's informatie, de A- en D- optimaliteits-

criteria worden toegepast en hun selectie mechanismen onderzocht. De testinformatiefuncties worden geëvalueerd als een functie van verdelingen van latente kenmerken en de itempool samenstelling. De afzonderlijke testsamenstellingen van de verschillende criteria biedt de mogelijkheid om het itemselectiecriterium af te stemmen op het specifieke doel van de test. De verkregen resultaten suggereren een grotere gevoeligheid van het D-optimaliteitscriterium voor verandering van de verdeling van de latente kenmerken. Het A-criterium lijkt daarentegen een duidelijke voorkeur voor zeer discriminative items te vertonen ongeacht de verdeling van de latente kenmerken. Wegens de gevoeligheid van het D-criterium voor de verdeling van latente kenmerken, wordt verwacht dat dit criterium geschikt is voor computerised adaptive testing (CAT).

Hoofdstuk vier evalueert de prestatie van het D-optimaliteitscriterium in CAT, en demonstreert de toepasbaarheid ervan. Twee andere criteria, het A-optimaliteit en het Kullback-Leibler (KL) informatie criterium, worden eveneens onderzocht. Een CAT simulatie studie is uitgevoerd waarbij nadruk gelegd wordt op de 'early stage' schattingen van latente kenmerken. De zogenaamde 'CAT early stage' wordt meestal gekarakteriseerd door instabiliteit van de parameterschattingen. Het is bekend dat in dit stadium de schattingen gecorrigeerd moeten worden voor bias. In het algemeen laat deze studie zien dat de spreiding van itembankinformatie over het continuüm van de latente kenmerken de voornaamste factor is in het bepalen van de nauwkeurigheid, precisie en stabiliteit van de schatting van latente kenmerken. Des te meer uniform deze spreiding van informatie is, des te groter is de kwaliteit van de schatting voor alle gebruikte criteria. Voor niet uniform verdeelde informatie presteren alle criteria ongeveer gelijk voor wat betreft meet-nauwkeurigheid en precisie aan het eind van de test. Ze verschillen echter in stabiliteit van de schattingen in de 'early stage'. Het A-criterium laat nadrukkelijke fluctuaties zien in de mean squared error (MSE) terwijl de D- en KL-criteria de attenuatie paradox effectiever lijken te omzeilen. Door de flexibiliteit van het D-optimaliteitscriterium bij het selecteren van items, waarbij afwisselend tussen hoge en lage discriminerende items wordt gekozen, lijkt het D-criterium een meer evenwichtig gebruik te maken van de beschikbare items in de pool.

Hoofdstuk vijf breidt dezelfde methoden van hoofdstuk vier uit naar een adaptive test met het GRM en geeft vergelijkbare resultaten m.b.t de robuustheid van het D-optimaliteitscriterium tegen instabiliteit in 'CAT early stage'. De specifieke selectie van itemparameter waarden door de criteria zijn nader onderzocht. Het probleem van 'item exposure' is ook geëvalueerd. Het resultaat dat het D-criterium een meer gebalanceerde gebruik van items in een pool maakt kon niet ten volle worden bevestigd. Wat 'item

exposure' betreft, hoe grote de itempool (met grotere verschillen tussen parameter waarden) des te meer overlap in exposure frequencies tussen de drie gebruikte criteria.

In de literatuur is weinig overeenstemming tussen opinies over welk soort items, polytome of dichotome, meer informatie over de onbekende latente kenmerken leveren. In het algemeen wordt aangenomen dat polytome items meer informatief zijn dan dichotome. Hoewel deze aanname waar is voor polytome items die door het GRM worden beschreven, is deze aanname niet altijd geldig zijn voor items met een nominale response format. Hoofdstuk zes laat voor het NRM zien dat, als men corrigeert voor de itemdiscriminatie parameters, de dichtome items even informatief kunnen zijn als hun polytome tegenhangers, zo niet meer. De bevindingen van test design optimalisatie waarin de items worden geselecteerd uit een samengestelde pool met polytome NRM en dichotome 2PL items, ondersteunt dit resultaat en laat zien dat het standpunt dat polytome items meer informatie leveren niet altijd gerechtvaardigd is. De testinformatie functie en de vorm van de test (paper&pencil tegenover adaptief) zijn de belangrijkste factoren bij de keuze tussen dichotome en polytome items. Er wordt geconcludeerd dat de winst in nauwkeurigheid die bereikt kan worden door polytome items, ingeruild zou kunnen worden voor de parametrische eenvoud van dichotome items.

Acknowledgment: I wish to gratefully acknowledge Marion de Leeuw's help, without which I would be literarily 'Lost in Translation'.